Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Decoding Imagined Speech from EEG using Transfer Learning

JERRIN, T P[1], (Student Member, IEEE) and RAMAKRISHNAN, A G [2], (Senior Member, IEEE)
[1]Department of Electrical Engineering, Indian Institute of Science, Bangalore, India (e-mail: jerrinp@iisc.ac.in)
[2]Department of Electrical Engineering, Indian Institute of Science, Bangalore, India (e-mail: agr@iisc.ac.in)

Corresponding author: Jerrin T P (e-mail: jerrinp@ iisc.ac.in).

The authors received no funding for this work.

ABSTRACT We present a transfer learning-based approach for decoding imagined speech from electroencephalogram (EEG). Features are extracted simultaneously from multiple EEG channels, rather than separately from individual channels. This helps in capturing the interrelationships between the cortical regions. To alleviate the problem of lack of enough data for training deep networks, sliding window-based data augmentation is performed. Mean phase coherence and magnitude-squared coherence, two popular measures used in EEG connectivity analysis, are used as features. These features are compactly arranged, exploiting their symmetry, to obtain a three dimensional "image-like" representation. The three dimensions of this matrix correspond to the alpha, beta and gamma EEG frequency bands. A deep network with ResNet50 as the base model is used for classifying the imagined prompts. The proposed method is tested on the publicly available ASU dataset of imagined speech EEG, comprising four different types of prompts. The accuracy of decoding the imagined prompt varies from a minimum of 79.7% for vowels to a maximum of 95.5% for short-long words across the various subjects. The accuracies obtained are better than the state-of-the-art methods, and the technique is good in decoding prompts of different complexities.

INDEX TERMS brain–computer interface, transfer learning, electroencephalogram, speech imagery, imagined speech

I. INTRODUCTION

Speech in both overt and covert forms are very natural to human beings since we learn to speak even without any formal education. During covert speech, we imagine speaking without any intentional movement of any of our articulators [1]. Decoding imagined speech from electroencephalogram (EEG) involves the discrimination between a fixed set of imagined words from the EEG captured during imagination. A system for decoding imagined speech has several applications including speech imagery BCI systems. In such BCI systems, speech imagery is used to generate distinct and repeatable neural activity. These systems can help patients whose muscles are paralyzed, as in the case of patients suffering from locked-in syndrome, to communicate with others and to operate devices such as computers [2].

Though it is almost a decade since the publication of the first research article on decoding imagined speech from EEG, the field has witnessed only slow progress compared to many other fields such as speech recognition [1]. This is primarily due to the lack of enough training data. Surplus training data is one of the primary reasons for the success of modern machine learning algorithms. This paper presents a novel deep learning architecture that addresses the scarcity of data in two ways: 1) by creating more data from the existing data using data augmentation and 2) by using transfer learning (TL) technique for training a deep network.

In most of the works in the literature such as [3]–[6], [8]–[18], [20]–[22], individual EEG channels are considered separately for extracting features. In these works except [22], [26], wavelet domain features, mel-frequency cepstral coefficients (MFCCs) and/or temporal domain features are extracted from each channel and are concatenated to obtain the feature vector for each trial. In [22], [26], the features extracted from individual channels are considered as distinct data vector and the decisions of the classifier for each channel are combined to obtain the final classification result. The feature extraction and classification techniques adopted in these works are comprehensively discussed in the review article [1]. Unlike these works where features are extracted from individual EEG channels, in this work, we extract

TABLE 1: Features and classifiers used in other works in the literature on decoding imagined speech from EEG. DNN: deep neural network, CNN: convolutional neural network, RNN: recurrent neural network, DAE: deep autoencoder, LSTM: long short-term memory, DWT: discrete wavelet transform, MFCC: mel frequency cepstral coefficients.

Sl. No.	Authors	**Features**	**Classifier**
1	Garcia et al. [3]	Discrete wavelet transform	Random forest
2	Brigham et al. [4]	Autoregressive model coefficients	Nearest neighbour
3	Min et al. [5]	Mean, variance, standard deviation, and skewness	Extreme learning machine
4	Sereshkeh et al. [6]	Discrete wavelet transform	Regularized neural network
5	Nguyen et al. [7]	Tangent vectors in Riemannian manifold	Relevance vector machine
6	Coonecy et al. [8]	MFCC, statistical features etc.	Support vector machine
7	Garcia et al. [9]	Bag of features	Naive Bayes
8	Tottrup et al. [10]	Spectral and temporal features	Random forest
9	Balaji et al. [11]	Spectral power	Artificial neural network
10	Jahangiri et al. [12]–[15]	Discrete Gabor transform	Linear discriminant analysis
11	Pawar and Dhage [16]	Discrete wavelet transform	Extreme learning machine
12	Koizumi et al. [17]	Spectral power	Support vector machine
13	Deng et al. [18]	Hilbert spectrum	Linear discriminant analysis
14	Zhang et al. [19]	Common spatial patterns	Support vector machine
15	Zhao and Rudzicz [20]	Statistical features	Support vector machine
16	Sereshkeh et al. [21]	Autoregressive model coefficients and DWT	Support vector machine
17	Panachakel et al. [22]	Temporal and DWT	Deep neural network
18	Saha et al. [23]	Channel cross-covariance	CNN + RNN + DAE
19	Saha et al. [24]	Channel cross-covariance	CNN + LSTM
20	Saha et al. [25]	Channel cross-covariance	CNN + DAE + XG Boost
21	Panachakel et al. [26]	Discrete wavelet transform	Deep neural network

the features simultaneously from multi-channel EEG. This approach is advantageous because studies have shown that complex cognitive tasks like speech production involve information transfer between multiple cortical areas. Extracting features separately from individual channels may not capture this interaction; however, when features are extracted simultaneously from multiple channels [27]–[30], this can be better captured.

A few works in the literature [7], [23], [24] have already employed simultaneous feature extraction from EEG for decoding imagined speech. These studies have used channel cross-covariance matrix for extracting the features. Features and classifiers used in popular works on decoding imagined speech from EEG are tabulated in Table 1.

Since it is difficult to have enough EEG data to train deep networks, a few researchers have recently taken recourse to transfer learning for decoding imagined speech from EEG. Transfer learning improves the performance of a classifier in the target domain by incorporating the knowledge gained from a different domain [9], [31], [32]. In the work by García-Salinas et al. [9], intra-subject transfer learning was applied to classify an imagined word using a classifier trained on a set of four words which does not include the target word. Cooney et al. [33] employed two different approaches for inter-subject transfer learning. A deep CNN architecture, similar to [34] is initially trained on a set of subjects called the source subjects and transfer learning is employed to improve the performance of the classifier on a new (target) subject.

In our work, EEG data is first augmented using sliding window method. From the augmented data, mean phase coherence (MPC) and magnitude-squared coherence (MSC) are extracted as features. A ResNet-50 [35] based transfer learning model is used as the classifier. We report the results of this approach on a publicly available dataset [7]. The dataset contains EEG recorded when the subjects were imagining four different types of prompts, namely vowels, short words, long words and short-long words. The proposed method achieves accuracies comparable to the state-of-the-

art results on the same dataset.

The major contributions of this work are listed below:

1) Although there are works in the literature where features are simultaneously extracted from multiple EEG channels, this is the first work to extract mean phase coherence and magnitude-squared coherence as features from EEG. Simultaneous extraction of features from multiple EEG channels helps in capturing the inter-electrode relationships related to speech imagery.
2) This is first work to make use of a deep network pre-trained for classifying images as the base classifier for classifying imagined speech EEG.
3) We exploit the symmetry in MPC and MSC for compactly packing them into an "image-like" 3-dimensional representation. The three dimensions correspond to three different EEG frequency bands.

The rest of the paper is organized as follows: Sec. II describes the dataset used in this work. Sec. III explains data augmentation using overlapping window. Sec. IV deals with the feature extraction method used in this work. The results we have obtained and the comparison with other works in the literature on the dataset we have used are given in Sec. V. Sec. VI enumerates the major limitations of this work.

II. DATASET USED FOR THE STUDY

The dataset used in this work was recorded by Human-Oriented Robotics and Controls (HORC) lab, Arizona State University and is publicly available for download [7]. The dataset consists of 64-channel EEG data acquired using BrainProducts ActiCHamp amplifier system. The four types of prompts used in the protocol are:

1) Long words: "independent" and "cooperate"
2) Short words: "in", "out" and "up"
3) Vowels: "/a/", "/i/" and "/u/"
4) Short-long words: "in" and "cooperate"

During the recording, the participants had to repeatedly imagine uttering the prompts without moving their articulators. The rate at which the prompt had to be imagined was cued using audio beeps. Although 15 subjects (S1 - S15, males = 11. females = 4) participated in the study, only the data of a subset of these participants is available for each protocol, details of which are listed in Table 2.

The data was recorded at a sampling rate of 1000 Hz but was later downsampled to 256 Hz. Also, a 5th order Butterworth bandpass filter with the pass band from 8 - 70 Hz was applied to remove any low frequency trends in the acquired signal and electromyogram (EMG) artifacts. A notch filter was used to remove the 60-Hz line noise. Ocular artifacts were removed using adaptive filtering [36]. More details about the dataset can be found in [7].

III. METHOD USED TO AUGMENT DATA

Data augmentation approaches such as overlapping or sliding window [37]–[40] and generative adversarial networks (GAN) [41]–[43] generate more training data from the existing data [44]. GAN is not the ideal approach for the current problem since the amount of available data is very limited. However, overlapping window can be used since the data consists of repeated imaginations in each trial. Accordingly, we have used overlapping windows as illustrated in Fig. 1. The length of the window is empirically chosen as 1 s (256 samples) and stride as 0.25 s (64 samples). This leads to an overlap of 0.75 s (192 samples). Using this approach, we can augment the data by a factor of 17.

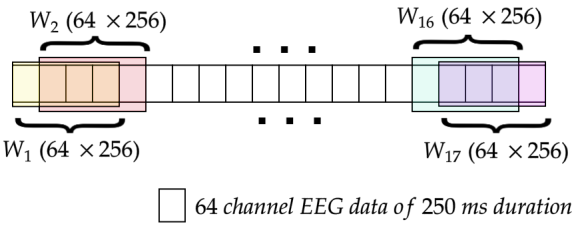


FIGURE 1: Data augmentation using overlapping windows. The initial EEG data has 1280 samples each from 64 channels (5 seconds of EEG at 256 Hz sampling rate). The window size is 256 samples or 1 s. The stride is 64 samples or 0.25 s which means an overlap of 75%. A total of 17 windows are obtained using this choice of window parameters.

We had attempted to use autocorrelation function to identify the repetitions in the EEG data and window the data based on the indices of repetition. However, it was difficult to detect the peaks in the autocorrelation function due to the low signal to noise ratio of the EEG signal.

IV. PARTICULARS OF FEATURE EXTRACTION

Unlike most works in the literature, we extract features simultaneously from all the EEG channels. Two measures are extracted as features:

1) Mean phase coherence
2) Magnitude-squared coherence

A. MEAN PHASE COHERENCE

Mean phase coherence (MPC) is a measure of phase synchronisation between two EEG channels [45]. MPC is closely related to phase locking value (PLV) defined for the condition where the phase difference between the studied channels is attributed to evoked activity [46]. PLV measures the phase synchronisation between two channels across different trials assuming that every trial is time-locked to a specific stimulus. This assumption does not hold good for EEG acquired during speech imagery since the imagination is not time-locked across trials albeit the presence of cues for the participant.

MPC across the i^{th} and k^{th} EEG channels is defined as,

$$MPC_{i,k} = \frac{1}{N}\left|\sum_{n=0}^{N-1} e^{-j(\phi_i(n)-\phi_k(n))}\right| \quad (1)$$

where N is the number of samples, $\phi_i(n)$ and $\phi_k(n)$ are the instantaneous phases of channels i and k at the n^{th} time sample. The instantaneous phases of channels are obtained

TABLE 2: Number of participants, whose data is available in each of the four protocols in the ASU imagined speech EEG dataset. Although 15 subjects participated in the study, only the data of a subset of them is available in the public dataset.

Protocol	Prompts	Total Number of Participants	IDs of the Participants
Long words	"independent" and "cooperate"	6	S2, S3, S6, S7, S9 and S11
Short words	"in", "out" and "up"	6	S1, S3, S5, S6, S8 and S12
Vowels	"/a/", "/i/" and "/u/"	8	S4, S5, S8, S9, S11, S12, S13 and S15
Short-long words	"in" and "cooperate"	6	S1, S5, S8, S9, S10 and S14

using Hilbert transform. The value of MPC lies between $[0, 1]$; a value close to zero indicates that the phase differences between the signals are random whereas a value of one means that the two signals are phase synchronized during most of the time interval considered [47]. MPC is used in epilepsy [48], [49] and sleep studies [50].

B. MAGNITUDE-SQUARED COHERENCE (MSC)

Coherence captures the linear relationship in the spectral domain [51]–[54] between a pair of signals. Let $S_{i,i}(\omega)$ and $S_{k,k}(\omega)$ denote the power spectral densities and $S_{i,k}(\omega)$ denote the cross power spectral density of X_{i*} and X_{k*}. The MSC between X_{i*} and X_{k*} is given by:

$$MSC_{i,k}(\omega) = \frac{|S_{i,k}(\omega)|^2}{S_{i,i}(\omega)S_{k,k}(\omega)} \tag{2}$$

The spectral densities are all estimated using Welch's overlapped averaged periodogram method [55]. Hamming window is used and the number of segments is eight. The values of MPC and MSC lie in the interval $[0, 1]$.

C. CONSTRUCTION OF INPUT AS 3D ARRAYS

Both MPC and MSC are frequency dependent measures. The input to the classifier are three-dimensional arrays with each dimension corresponding to one of the frequency bands, alpha (8 to 13 Hz), beta (13 to 30 Hz) and gamma (30 to 70 Hz). The EEG data is bandpass filtered to obtain the MPC of each EEG band. The MSC matrices for all the frequencies in a given band are averaged. The bands below 8 Hz and above 70 Hz are not used since the publicly available dataset is band pass filtered between 8 and 70 Hz.

There are six matrices corresponding to the three frequency bands and two measures (MPC and MSC). Since the input to the classifier needs to be similar to the images in the ImageNet [56] database which our classifier is pretrained on, we have compactly arranged the 6 matrices into a three-dimensional array. In the case of a regular RGB image (such as the images in ImageNet), each parallel plane corresponds to one of the three colours: red, green and blue. In our case, each parallel plane in the input three-dimensional arrays corresponds to one of the three frequency bands: alpha, beta and gamma. The compact arrangement is possible because of the symmetry of the matrices. Since both MPC and MSC matrices are symmetric, no information is lost if the upper or lower triangular elements of the matrices are removed. Therefore, a new matrix is created, consisting of the upper triangular elements of MPC and lower triangular elements of MSC of each band. Each of these newly constructed matrices corresponds to one of the three frequency bands. The diagonal elements of all the matrices are made zero and these matrices are combined to form the three-dimensional input array. Thus we have the information from both MPC and MSC across the three bands compactly placed in a three-dimensional array. If I is one of these three-dimensional arrays, then $I(i, j, 1)$, $I(i, j, 2)$ and $I(i, j, 3)$ respectively denote the alpha, beta, and gamma band information. Further, $I(i, j, 1)$ for $i > j$ denotes the MSC values in the alpha band whereas $I(i, j, 1)$ for $i < j$ denotes the MPC values in the alpha band. $I(i, j, :)$ for $i = j$ is zero for all the three bands. Figure 2 shows the various steps in generating the input to the classifier.

D. DETAILS OF THE CLASSIFIER

The architecture of the network used in this work is shown in Fig. 3. A ResNet50 [35] based deep neural network model, pre-trained on ImageNet [56], is used as the base model for the classifier. Since this network is trained to classify 1000 object categories, the output layer is a fully connected (FC) layer with 1000 neurons and softmax activation function. We replaced this layer with two FC layers with ReLU activation function and one output layer with softmax activation function. The first and second FC layers (FC1, FC2) have 128 and 64 neurons, respectively. The number of neurons in the output layer is the same as the number of classes in the imagined prompt category. For long and long-short words, the number of classes is two whereas for short words and vowels, it is three.

During the training on imagined speech data, the ResNet model layers are frozen and only the appended FC layers are trained. Thus the ResNet layers act as feature extraction layers. Adam optimizer [57] is used with cross-entropy loss function and a learning rate of $1e{-}4$. 10-fold cross-validation is performed on the data of each subject. The data is divided into 10 folds and during each cross-validation iteration, 9 out of the 10 folds are used for training and the remaining fold, for testing. This is repeated 10 times so that all the folds are

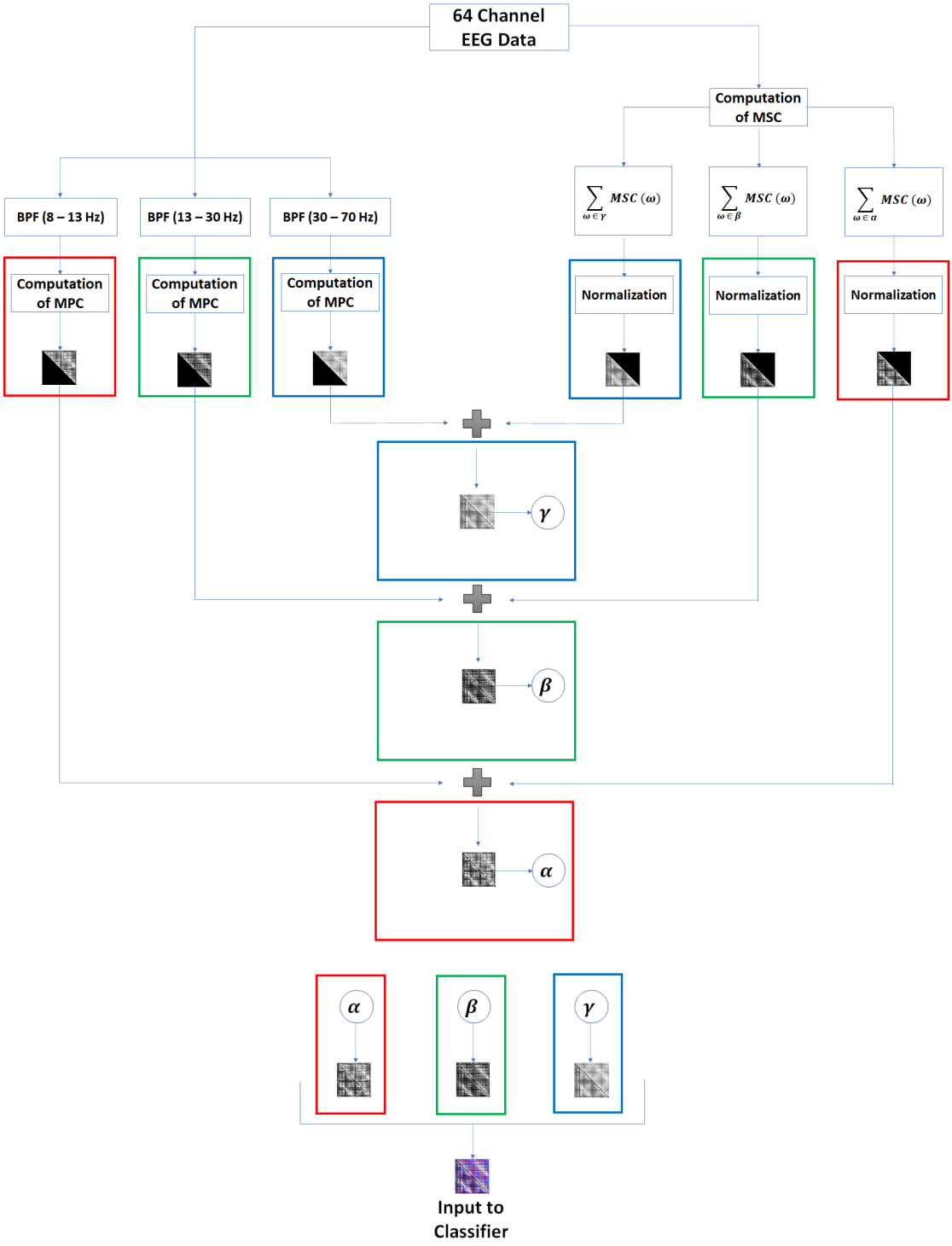


FIGURE 2: Illustration of various steps in the creation of the input to the classifier. In the case of training data, we start with 64×256 EEG data whereas with test data, we start with 64×1280. This difference is because of the fact that data augmentation is performed only for the training data and not for the test data. **BPF** denotes bandpass filter. Outline boxes of red, green and blue colours denote the data in the alpha, beta and gamma bands, respectively. $\sum_{\omega \in \alpha}$, $\sum_{\omega \in \beta}$ and $\sum_{\omega \in \gamma}$ respectively denote the summation of all the MSC matrices corresponding to alpha, beta and gamma bands. The dimension of the arrays that are input to the classifier is $64 \times 64 \times 3$.

tested once. Only the training data is augmented. This does not affect the dimensionality of the input data during testing since the latter is determined by the number of channels and does not vary with the length of the EEG signal.

During each cross-validation step, 85% of the training data is used for training the classifier and the remaining 15%, for validation. The maximum number of epochs for training is 100. To avoid overfitting, we have also implemented early stopping based on validation accuracy with the patience parameter set to 30 epochs. We ran the code on Google Colab. The GPU configuration in the session was NVIDIA Tesla T4 with GDDR6 RAM. The approximate time for training the network excluding feature extraction, file reading and other overheads was around 4 minutes.

V. RESULTS OF OUR STUDY

We tested the proposed methods on all the types of speech imagery available in the dataset. The accuracies obtained by the classifier for different classes of imagined prompts are listed in Tables 3, 4, 5 and 6, and compared with the state-of-the-art results in the literature. Clearly, the accuracy for every subject is better than the best in the literature for every type of prompt. The accuracy of decoding the imagined prompt varies from a minimum of 79.7% for vowels for the subject S13 to a maximum of 95.5% for short-long words for the subject S1. Figure 4 compares the mean accuracy across the subjects for each class of imagined prompts with the best two techniques reported in the literature.

To understand the effectiveness of data augmentation, we trained the classifier separately using the actual and the augmented data. Figure 5 compares the performance of our technique with and without data augmentation for the task of classifying short-long words. With data augmentation, the accuracy for all the subjects is above 90%, with the maximum of 95.5% for S1. Clearly, the accuracy of the system drops to mere chance level performance when the data augmentation stage is removed. This is expected due to the reduction in the number of training data.

Figure 6 gives the overall confusion matrices for each of the four classes of prompts. The precision and recall of various prompts of all the four classes of prompts are listed in Table 7. These values are consistently high, ranging from about 84% for short words to about 92.5% for short-long words. The precision and recall of all the prompts within each class of prompts are comparable indicating that the classifiers are not favouring any particular prompt within any class. The proposed method has performance comparable to the method developed by Saha and Fels [23] where a combination of CNN and recurrent neural networks (RNN) were used. Although RNNs are capable of capturing time-series information, they failed to give performance superior to our results in our experiments with the same feature extraction techniques. This might be because of the sliding window data augmentation applied in this work, which disturbs the time-series information in the data.

Since the number of classes in different types of imagined prompts are different, we have also computed Cohen's kappa (κ) value for the different classifiers. Kappa is defined as:

$$\kappa := \frac{p_{cl} - p_{ch}}{100 - p_{ch}} \tag{3}$$

where p_{cl} is the accuracy of the classifier and p_{ch} is the chance level accuracy, both in percentage. The value of κ lies in the range $[-1, 1]$. Values closer to 0 indicate that the classifier is only as good as random guess whereas values less than 0 indicate that the performance is inferior to random guess. κ values of various classifications are compared with those of the best techniques in the literature in Figs. 7, 8, 9, and 10 for long-words, short-words, vowels, and short-long words, respectively. The mean values of κ for long words, short words, vowels, and short-long words for the proposed method are 0.78 ± 0.01, 0.75 ± 0.05, 0.78 ± 0.06 and 0.87 ± 0.02, respectively. Clearly, short-long words have the highest mean and the lowest standard deviation among all the types of prompts. This may be because of the difference between the complexity of the two words, "in" and "cooperate" ("in" is monosyllabic containing a nasal consonant whereas "cooperate" is quadrisyllabic with no nasals). This result is in-line with the observation of Nguyen et al. in [7].

Further, to study the effectiveness of the individual EEG frequency bands in decoding speech imagery, we subdivided each of them into three subbands. The subbands within each frequency band are chosen such that they have approximately equal bandwidths in the logarithmic frequency scale. Table 8 lists the subbands chosen within the alpha, beta and gamma bands. Decoding experiments were separately conducted using only the features extracted from the subbands of each of the main bands. The results of using these subbands for classifying short-long words are shown in Fig. 11, along with the results obtained using all the three undivided EEG bands together. Clearly, the alpha and gamma bands have the lowest and the highest accuracies, respectively. This is in-line with the observation of Koizumi et al. [17]. Gamma band gives a performance which is nearly 20% higher than that of alpha band. For subjects S8 and S10, the accuracies of beta band and gamma band are comparable. This trend was observed in one of our previous works on classifying imagined phonemes [58] where a different dataset was used. The accuracy obtained by combining all the three bands is higher than the accuracy with gamma band by a minimum of 5% and a maximum of 14% for different subjects, clearly indicating the need for using all the three EEG frequency bands to obtain the best performance.

VI. LIMITATIONS OF THE WORK

The following are the limitations of the current work:

- As with many works employing deep learning, we cannot pinpoint to a particular part of the feature vector that leads to the good performance of the system. It would have been better if we are able to pinpoint a subset of EEG channels that have good discriminatory power for the imagined prompts. In our previous work [26],

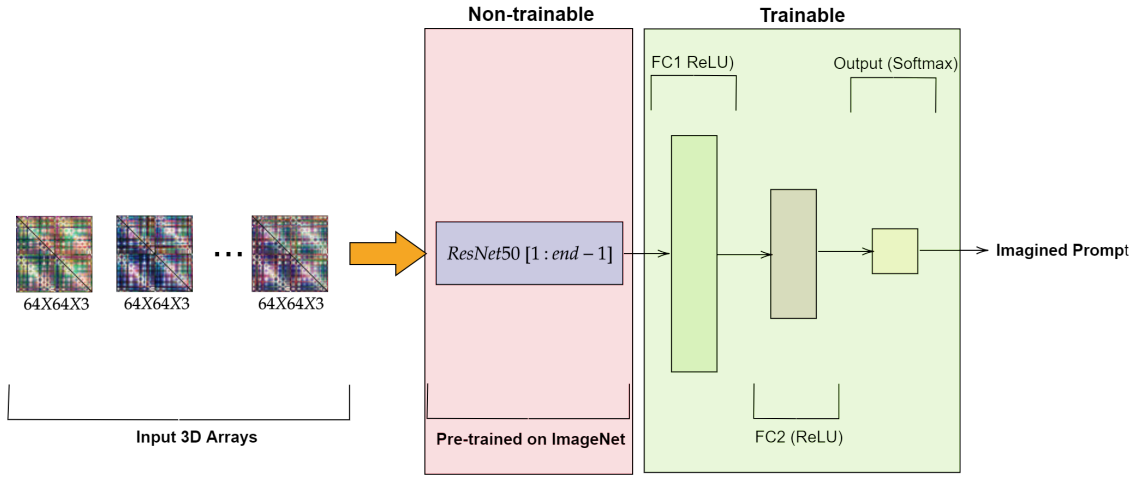


FIGURE 3: Architecture of the proposed deep network for decoding imagined speech. FC denotes a fully connected layer. The dimension of the input data arrays is $64 \times 64 \times 3$. FC1 and FC2 have 128 and 64 neurons, respectively. The number of neurons in the output layer is the same as the number of classes in each category of imagined prompts. Layers with 30% dropout after each FC layer are not shown. $ResNet[1 : end - 1]$ denotes the ResNet model pre-trained on ImageNet with the last FC layer removed. The parameters in the ResNet model are frozen and only the FC layers after the ResNet model are trained using the imagined speech EEG.

TABLE 3: Comparison of our accuracies with similar studies on classifying long words, viz. "independent" and "cooperate". The accuracies in percentage are shown in the format mean ± std. dev. Standard deviation values are not reported in [23].

Method/Participant ID	S2	S3	S6	S7	S9	S11
Tangent + RVM [7]	70.8±7.8	64.3±6.6	72.0±0.6	64.5±5.5	67.8±6.8	58.5±7.4
Hierarchical deep features [23]	77.5	90.7	73.7	86.8	80.1	71.1
Proposed (ResNet50+TL)	91.8±5.3	91.5±4.4	85.5±4.1	92.5±6.7	88.3±6.6	83.3±3.4

TABLE 4: Comparison of our accuracies with similar studies on classifying short words, viz. "in", "out" and "up". The accuracies in percentage are shown in the format mean ± std. dev. Standard deviation values are not reported in [23].

Method/Participant ID	S1	S3	S5	S6	S8	S12
Tangent + RVM [7]	48.0 ± 6.1	49.7 ± 5.5	46.3 ± 8.2	54.0 ± 9.1	47.7 ± 9.8	54.7 ± 6.9
Hierarchical deep features [23]	69	85	71	68	72	77
Proposed (ResNet50+TL)	81.7± 6.2	85.7 ± 5.8	80.0 ± 5.2	83.3 ± 5.9	84.3 ± 4.2	88.7 ± 4.6

TABLE 5: Comparison of our accuracies with similar studies on classifying the vowels "/a/", "/i/" and "/u/". The accuracies in percentage are shown in the format mean ± std. dev. Standard deviation values are not reported in [23].

Method/Participant ID	S4	S5	S8	S9	S11	S12	S13	S15
Tangent + RVM [7]	47.0 ±4.6	48.0 ±7.2	51.0 ±6.7	47.0 ±5.5	53.0 ±4.0	51.0 ±6.3	46.7 ±8.2	48.0 ±7.2
Hierarchical deep features [23]	69	85	71	68	72	77	69	83
Proposed (ResNet50+TL)	83.3 ±5.3	93.3 ±2.3	89.9 ±5.5	85.33 ±6.2	87.7 ±4.2	81.3 ±4.9	79.7 ±4.2	89.7 ±5.7

TABLE 6: Comparison of our accuracies with similar studies on classifying short-long words, viz. “in” and “cooperate”. The accuracies in percentage are shown in the format mean ± std. dev.

Method/Participant ID	S1	S5	S8	S9	S10	S14
Tangent + RVM [7]	70.3 ± 5.5	71.5 ± 5.0	81.9 ± 6.5	88.0 ± 6.4	79.3 ± 7.7	89.3 ± 3.5
Wavelet + DNN [22]	65.5 ± 9.6	64.5 ± 10.3	71.0 ± 5.3	86.2 ± 8.7	76.3 ± 5.6	77.2 ± 5.3
Proposed (ResNet50+TL)	95.5 ± 4.3	94.0 ± 2.1	92.5 ± 1.7	91.4 ± 4.5	90.1 ± 3.7	93.3 ± 2.9

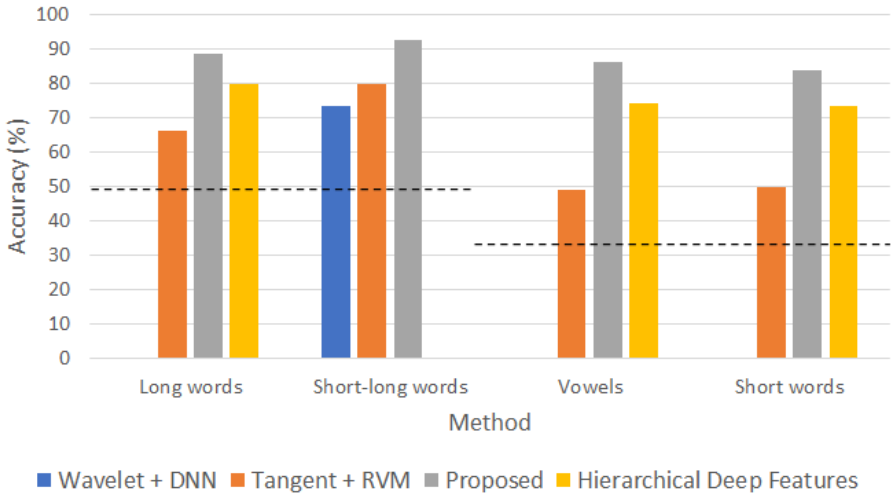


FIGURE 4: Performance comparison of the proposed (ResNet50+TL) technique with similar studies in the literature in terms of the mean accuracy of classification of each type of prompts: long words, short-long words, vowels and short words. The chance level accuracy for each type of prompts is shown by the dashed line.

we have used common spatial patterns to identify the EEG channels of interest. We have also shown how the accuracy varies with the number of channels used for feature extraction. A similar analysis is difficult here due to the huge computational cost and complexity of the architecture. However, this high computational cost is associated only with the training and not with testing.

- A similar analysis in terms of the prompts would also be interesting. As we have shown in Sec. V, length of the prompts do have good discriminatory powers. In one of our other works, we have shown that MPC values can be used for discriminating imagined prompts at the phonological level [58]. A deeper analysis into what the features are actually capturing can help in designing better prompts for speech imagery based BCI systems.
- We have used two frequency dependent measures popular in connectivity analysis. Further analysis is required to ascertain the differences in the information captured by the two measures. The two measures are not exactly the same, since if they were, all the input matrices would have been perfectly symmetric. However, some similarities are observed between the lower and upper triangular entries of the input matrices in the beta band. The reason for this similarity in some regions and lack of it in other regions is an interesting topic for investigation, considering the difference between the mathematical formulations of the two measures.
- It would be interesting to quantify the effect of the length of window and the stride used for data augmentation on the classification accuracy.

VII. CONCLUSION

This work proposes a novel transfer learning based architecture for decoding imagined speech. The training data is augmented using overlapping analysis windows to alleviate the problem of having limited training data. A deep network with ResNet50 network pre-trained on ImageNet as the base classifier is used for classifying the imagined prompts. The results obtained are superior to the state-of-the-art results for every type of prompt and on every subject for the ASU dataset used. This is the first work to use a network trained for classifying real-world images for classifying imagined prompts. Unlike most works in the literature on classifying imagined prompts, we extract features simultaneously from multiple EEG channels. This helps us to capture the inter-channel interactions involved in speech imagery. Two popular measures used in EEG connectivity analysis, namely mean phase coherence and magnitude-squared coherence are

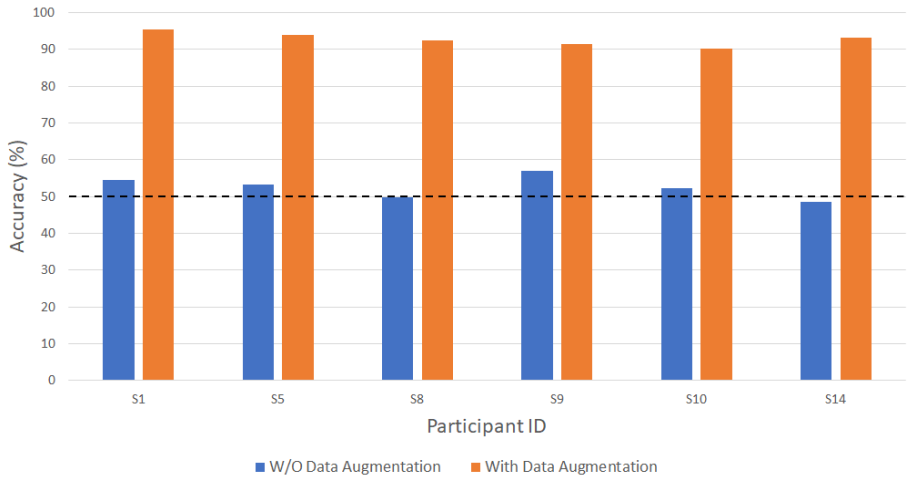


FIGURE 5: Comparison of the performance of the proposed (ResNet50+TL) technique with and without data augmentation, in terms of the mean accuracy of classifying the short-long words "in" and "cooperate". The data augmentation increases the accuracy by about 40% in all the cases. The chance level accuracy is shown by the dashed line.

TABLE 7: Table showing the precision and recall of all the prompts of various classes of prompts considered independently. TP: number of true positives, FP: number of false positives, TN: number of true negatives and FN: number of false negatives

Prompt	TP	FP	TN	FN	Precision (%)	Recall (%)
Long words						
independent	532	66	534	68	89.0	88.7
cooperate	534	68	532	66	88.7	89.0
Short-long words						
in	518	39	521	42	93.0	92.5
cooperate	521	42	518	39	92.5	93.0
Short words						
out	1005	194	2206	195	83.8	83.8
in	1009	191	2209	191	84.1	84.1
up	1008	193	2207	192	83.9	84.0
Vowels						
/a/	1378	216	2984	222	86.4	86.1
/i/	1377	225	2975	223	86.0	86.1
/u/	1380	224	2976	220	86.0	86.3

TABLE 8: Subbands of EEG frequency bands used in this study. The subbands are chosen such that bandwidths within each band are approximately equal in the logarithmic frequency scale.

EEG Band	Subband 1	Subband 2	Subband 3
Alpha	8 - 9.5 Hz	9.5 - 11 Hz	11 - 13 Hz
Beta	13 - 17.2 Hz	17.2 - 22.7 Hz	22.7 - 30 Hz
Gamma	30 - 39.8 Hz	39.8 - 52.8 Hz	52.8 - 70 Hz

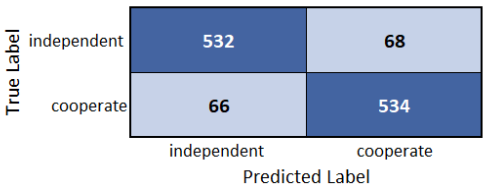
used as features. These measures are compactly packed into a three dimensional array, resembling the images in ImageNet used for pre-training ResNet50. The compact packing reduces the dimensions of the input to the classifier.

ACKNOWLEDGEMENT

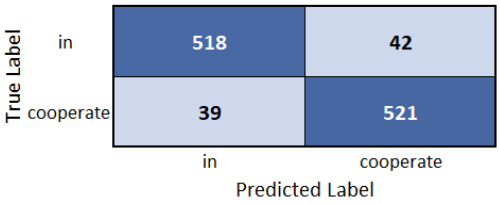
We thank Mr. Pradeep Kumar G. and Mr. Anoop C.S., MILE Lab, Indian Institute of Science, Bangalore for the support extended to this work.

REFERENCES

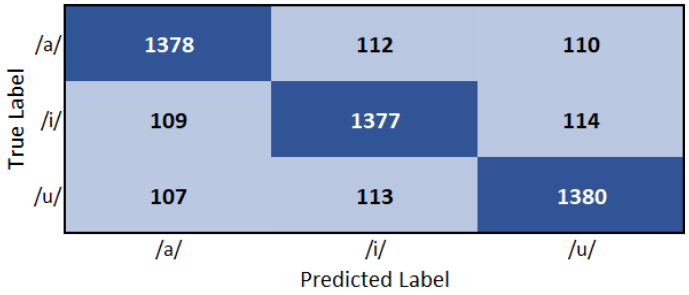
[1] J. T. Panachakel and R. A. Ganesan, "Decoding covert speech from EEG-a comprehensive review," Frontiers in Neuroscience, vol. 15, p. 392, 2021.
[2] S. N. Abdulkader, A. Atia, and M.-S. M. Mostafa, "Brain computer interfacing: Applications and challenges," Egyptian Informatics Journal, vol. 16, no. 2, pp. 213–230, 2015.
[3] A. A. T. García, C. A. R. García, and L. V. Pineda, "Toward a silent speech interface based on unspoken speech." in BIOSIGNALS, 2012, pp. 370–373.
[4] K. Brigham and B. V. Kumar, "Imagined speech classification with EEG signals for silent communication: a preliminary investigation into synthetic telepathy," in Bioinformatics and Biomedical Engineering (iCBBE), 2010 4th International Conference on. IEEE, 2010, pp. 1–4.
[5] B. Min, J. Kim, H.-j. Park, and B. Lee, "Vowel imagery decoding toward silent speech BCI using extreme learning machine with electroencephalogram," BioMed research international, vol. 2016, 2016.



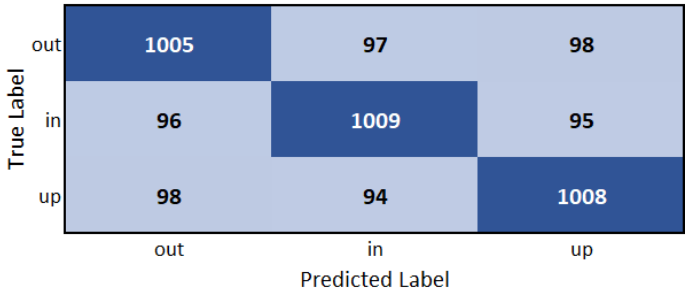
(a) Long words. For prompt "independent", $TPR = 88.6\%$ and $TNR = 89.0\%$; for prompt "cooperate", $TPR = 89.0\%$ and $TNR = 88.6\%$



(b) Short-long words. For prompt "in", $TPR = 92.5\%$ and $TNR = 93.0\%$; for prompt "cooperate", $TPR = 93.0\%$ and $TNR = 92.5\%$



(c) Vowels. For prompt "/a/", $TPR = 86.1\%$ and $TNR = 86.2\%$; for prompt "/i/", $TPR = 86.1\%$ and $TNR = 86.2\%$; for prompt "/u/", $TPR = 86.3\%$ and $TNR = 86.1\%$



(d) Short words. For prompt "out", $TPR = 83.8\%$ and $TNR = 84.0\%$; for prompt "in", $TPR = 84.1\%$ and $TNR = 84.0\%$; for prompt "up", $TPR = 84.0\%$ and $TNR = 83.9\%$

FIGURE 6: Confusion matrices for each type of prompts across all the participants. The number of trials per prompt for long words, short-long words, vowels and short words are 600, 560, 1200, and 1600 respectively. TPR : True positive rate, TNR : True negative rate

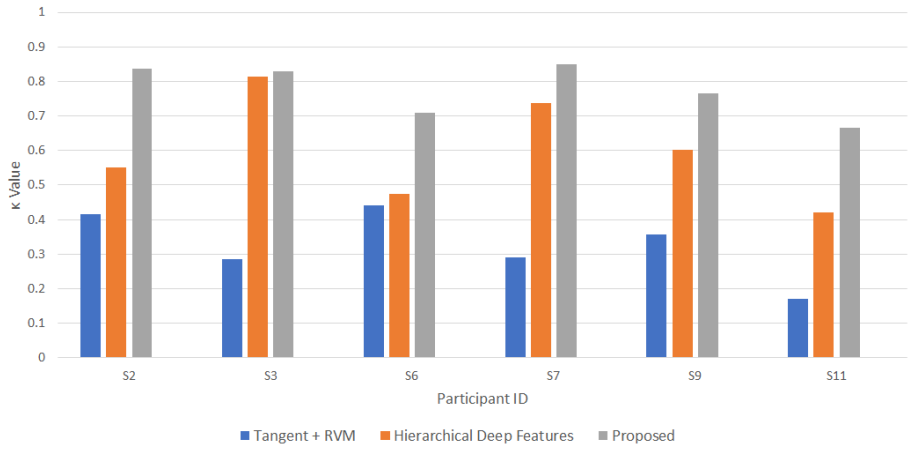


FIGURE 7: Comparison of κ values with similar studies on classifying long words, viz. “independent” and “cooperate”.

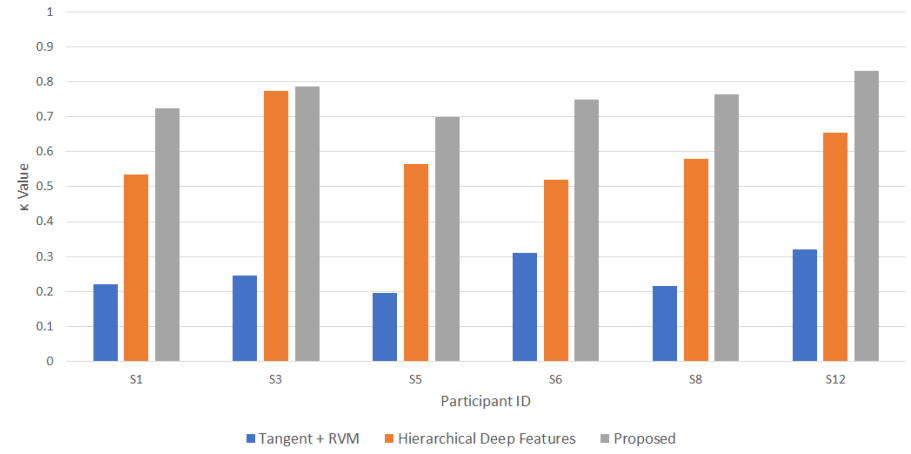


FIGURE 8: Comparison of κ values with similar studies on classifying short words, viz. “in”, “out” and “up”.

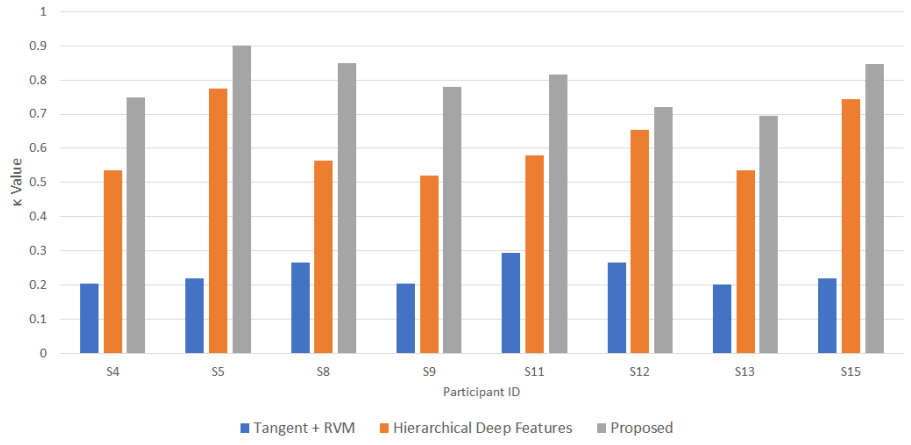


FIGURE 9: Comparison of κ values with similar studies on classifying the vowels “/a/”, “/i/” and “/u/”.

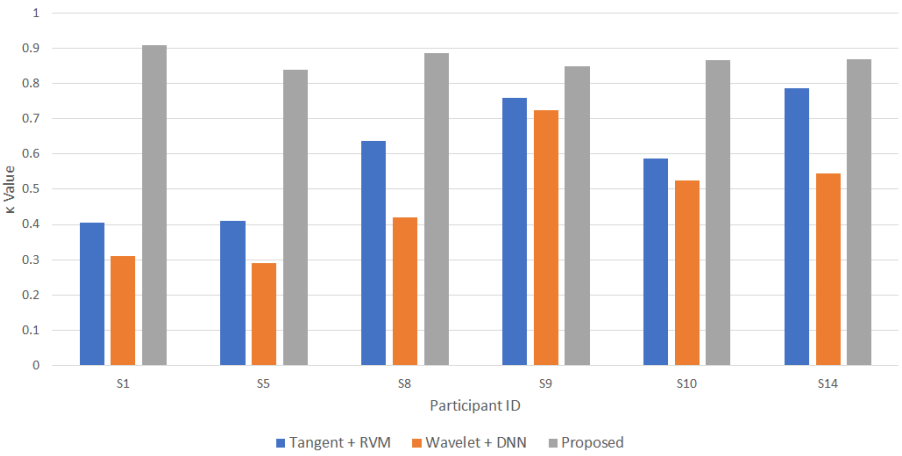


FIGURE 10: Comparison of κ values with similar studies on classifying short-long words, viz. "in" and "cooperate".

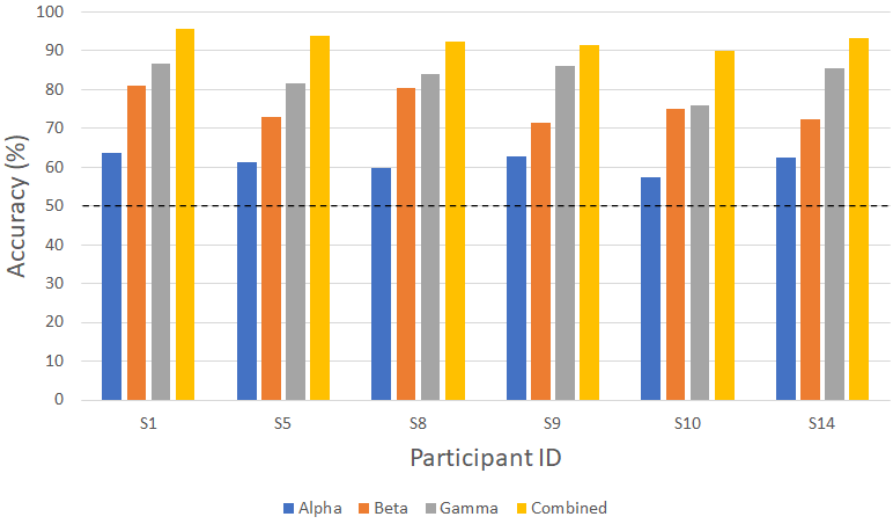


FIGURE 11: Performance comparison of the proposed (ResNet50+TL) approach when subbands within each EEG frequency band, alpha, beta and gamma are used to train the classifier for classifying short-long words, namely "in" and "cooperate". Each frequency band is further divided into three subbands which are equally spaced in the logarithmic scale. Gamma band consistently outperforms the other bands and the accuracy significantly improves when all the bands are combined. The chance level accuracy is shown by the dashed line.

[6] A. R. Sereshkeh, R. Trott, A. Bricout, and T. Chau, "EEG classification of covert speech using regularized neural networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 12, pp. 2292–2300, 2017.
[7] C. H. Nguyen, G. K. Karavas, and P. Artemiadis, "Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features," Journal of neural engineering, vol. 15, no. 1, p. 016002, 2017.
[8] C. Cooney, R. Folli, and D. Coyle, "Mel frequency cepstral coefficients enhance imagined speech decoding accuracy from EEG," in 2018 29th Irish Signals and Systems Conference (ISSC). IEEE, 2018, pp. 1–7.
[9] J. S. García-Salinas, L. Villaseñor-Pineda, C. A. Reyes-García, and A. A. Torres-García, "Transfer learning in imagined speech EEG-based BCIs," Biomedical Signal Processing and Control, vol. 50, pp. 151–157, 2019.
[10] L. Tøttrup, K. Leerskov, J. T. Hadsund, E. N. Kamavuako, R. L. Kæseler, and M. Jochumsen, "Decoding covert speech for intuitive control of brain-computer interfaces based on single-trial EEG: a feasibility study," in 2019 IEEE 16th International Conference on Rehabilitation Robotics (ICORR). IEEE, 2019, pp. 689–693.
[11] A. Balaji, A. Haldar, K. Patil, T. S. Ruthvik, C. Valliappan, M. Jartarkar, and V. Baths, "EEG-based classification of bilingual unspoken speech using ANN," in Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE. IEEE, 2017, pp. 1022–1025.
[12] A. Jahangiri, D. Achanccaray, and F. Sepulveda, "A novel EEG-based four-class linguistic BCI," in 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2019, pp. 3050–3053.
[13] A. Jahangiri, J. M. Chau, D. R. Achanccaray, and F. Sepulveda, "Covert speech vs. motor imagery: a comparative study of class separability in identical environments," in 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2018, pp. 2020–2023.
[14] A. Jahangiri and F. Sepulveda, "The contribution of different frequency bands in class separability of covert speech tasks for BCIs," in 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2017, pp. 2093–2096.
[15] ——, "The relative contribution of high-gamma linguistic processing stages of word production, and motor imagery of articulation in class separability of covert speech tasks in EEG data," Journal of medical systems, vol. 43, no. 2, p. 20, 2019.

[16] D. Pawar and S. Dhage, "Multiclass covert speech classification using extreme learning machine," Biomedical Engineering Letters, pp. 1–10, 2020.

[17] K. Koizumi, K. Ueda, and M. Nakao, "Development of a cognitive brain-machine interface based on a visual imagery method," in 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2018, pp. 1062–1065.

[18] S. Deng, R. Srinivasan, T. Lappas, and M. D'Zmura, "EEG classification of imagined syllable rhythm using hilbert spectrum methods," Journal of neural engineering, vol. 7, no. 4, p. 046006, 2010.

[19] X. Zhang, H. Li, and F. Chen, "EEG-based classification of imaginary Mandarin tones," in 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2020, pp. 3889–3892.

[20] S. Zhao and F. Rudzicz, "Classifying phonological categories in imagined and articulated speech," in Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE, 2015, pp. 992–996.

[21] A. R. Sereshkeh, R. Trott, A. Bricout, and T. Chau, "Online EEG classification of covert speech for brain–computer interfacing," International journal of neural systems, vol. 27, no. 08, p. 1750033, 2017.

[22] J. T. Panachakel, A. Ramakrishnan, and T. Ananthapadmanabha, "A novel deep learning architecture for decoding imagined speech from EEG," arXiv preprint arXiv:2003.09374, 2020.

[23] P. Saha and S. Fels, "Hierarchical deep feature learning for decoding imagined speech from EEG," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 10 019–10 020.

[24] P. Saha, S. Fels, and M. Abdul-Mageed, "Deep learning the EEG manifold for phonological categorization from active thoughts," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 2762–2766.

[25] P. Saha, M. Abdul-Mageed, and S. Fels, "Speak your mind! towards imagined speech recognition with hierarchical deep learning," arXiv preprint arXiv:1904.05746, 2019.

[26] J. T. Panachakel, A. Ramakrishnan, and T. Ananthapadmanabha, "Decoding imagined speech using wavelet features and deep neural networks," in 2019 IEEE 16th India Council International Conference (INDICON). IEEE, 2019, pp. 1–4.

[27] J. M. Correia, C. Caballero-Gaudes, S. Guediche, and M. Carreiras, "Phonatory and articulatory representations of speech production in cortical and subcortical fMRI responses," Scientific Reports, vol. 10, no. 1, pp. 1–14, 2020.

[28] K. Simonyan and S. Fuertinger, "Speech networks at rest and in action: interactions between functional brain networks controlling speech production," Journal of neurophysiology, vol. 113, no. 7, pp. 2967–2978, 2015.

[29] S. Fuertinger, B. Horwitz, and K. Simonyan, "The functional connectome of speech control," PLoS Biol, vol. 13, no. 7, p. e1002209, 2015.

[30] S.-H. Lee, M. Lee, and S.-W. Lee, "Functional connectivity of imagined speech and visual imagery based on spectral dynamics," arXiv preprint arXiv:2012.03520, 2020.

[31] S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Transactions on knowledge and data engineering, vol. 22, no. 10, pp. 1345–1359, 2009.

[32] H. He and D. Wu, "Transfer learning enhanced common spatial pattern filtering for brain computer interfaces (BCIs): Overview and a new approach," in International Conference on Neural Information Processing. Springer, 2017, pp. 811–821.

[33] C. Cooney, R. Folli, and D. Coyle, "Optimizing layers improves CNN generalization and transfer learning for imagined speech decoding from EEG," in 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC). IEEE, 2019, pp. 1311–1316.

[34] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," Human brain mapping, vol. 38, no. 11, pp. 5391–5420, 2017.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[36] P. He, G. Wilson, and C. Russell, "Removal of ocular artifacts from electro-encephalogram by adaptive filtering," Medical and biological engineering and computing, vol. 42, no. 3, pp. 407–412, 2004.

[37] A. O'Shea, G. Lightbody, G. Boylan, and A. Temko, "Neonatal seizure detection using convolutional neural networks," in 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2017, pp. 1–6.

[38] N.-S. Kwak, K.-R. Müller, and S.-W. Lee, "A convolutional neural network for steady state visual evoked potential classification under ambulatory environment," PloS one, vol. 12, no. 2, p. e0172578, 2017.

[39] I. Ullah, M. Hussain, H. Aboalsamh et al., "An automated system for epilepsy detection using EEG brain signals based on deep learning approach," Expert Systems with Applications, vol. 107, pp. 61–71, 2018.

[40] I. Majidov and T. Whangbo, "Efficient classification of motor imagery electroencephalography signals using deep learning methods," Sensors, vol. 19, no. 7, p. 1736, 2019.

[41] Y. Luo and B.-L. Lu, "EEG data augmentation for emotion recognition using a conditional Wasserstein GAN," in 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2018, pp. 2535–2538.

[42] Z. Wei, J. Zou, J. Zhang, and J. Xu, "Automatic epileptic EEG detection using convolutional neural network with improvements in time-domain," Biomedical Signal Processing and Control, vol. 53, p. 101551, 2019.

[43] S. Chang and H. Jun, "Hybrid deep-learning model to recognise emotional responses of users towards architectural design alternatives," Journal of Asian Architecture and Building Engineering, vol. 18, no. 5, pp. 381–391, 2019.

[44] E. Lashgari, D. Liang, and U. Maoz, "Data augmentation for deep-learning-based electroencephalography," Journal of Neuroscience Methods, p. 108885, 2020.

[45] F. Mormann, K. Lehnertz, P. David, and C. E. Elger, "Mean phase coherence as a measure for phase synchronization and its application to the EEG of epilepsy patients," Physica D: Nonlinear Phenomena, vol. 144, no. 3-4, pp. 358–369, 2000.

[46] R. Bruña, F. Maestú, and E. Pereda, "Phase locking value revisited: teaching new tricks to an old dog," Journal of neural engineering, vol. 15, no. 5, p. 056011, 2018.

[47] R. Q. Quiroga and S. Panzeri, Principles of neural coding. CRC Press, 2013.

[48] H. S. Alaei, M. A. Khalilzadeh, and A. Gorji, "Optimal selection of SOP and SPH using fuzzy inference system for on-line epileptic seizure prediction based on EEG phase synchronization," Australasian physical & engineering sciences in medicine, vol. 42, no. 4, pp. 1049–1068, 2019.

[49] F. Mormann, R. G. Andrzejak, T. Kreuz, C. Rieke, P. David, C. E. Elger, and K. Lehnertz, "Automated detection of a preseizure state based on a decrease in synchronization in intracranial electroencephalogram recordings from epilepsy patients," Physical Review E, vol. 67, no. 2, p. 021912, 2003.

[50] K. Mezeiová and M. Paluš, "Comparison of coherence and phase synchronization of the human sleep electroencephalogram," Clinical Neurophysiology, vol. 123, no. 9, pp. 1821–1830, 2012.

[51] A. M. Bastos and J.-M. Schoffelen, "A tutorial review of functional connectivity analysis methods and their interpretational pitfalls," Frontiers in systems neuroscience, vol. 9, p. 175, 2016.

[52] A. H. Ghaderi, S. Moradkhani, A. Haghighatfard, F. Akrami, Z. Khayyer, and F. Balcı, "Time estimation and beta segregation: An EEG study and graph theoretical approach," PLoS One, vol. 13, no. 4, p. e0195380, 2018.

[53] A. H. Ghaderi, M. A. Nazari, H. Shahrokhi, and A. H. Darooneh, "Functional brain connectivity differences between different ADHD presentations: Impaired functional segregation in ADHD-combined presentation but not in ADHD-inattentive presentation," Basic and Clinical Neuroscience, vol. 8, no. 4, p. 267, 2017.

[54] S. M. Kay, Modern spectral estimation: theory and application. Pearson Education India, 1988.

[55] P. Welch, "The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms," IEEE Transactions on audio and electroacoustics, vol. 15, no. 2, pp. 70–73, 1967.

[56] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255.

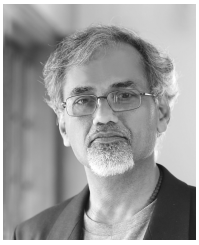
[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[58] J. T. Panachakel and A. Ramakrishnan, "Classification of phonological categories in imagined speech using phase synchronization measure," in 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2021.

JERRIN THOMAS PANACHAKEL (M'20) received his B.Tech. and M.Tech. degrees in Electronics and Communication Engineering in the year 2012 and 2014 from the University of Kerala, India. He is currently pursuing his Ph.D. in Electrical Engineering at Indian Institute of Science, Bangalore, India.

From 2014 to 2015, he was a Research Associate at MILE Lab, Indian Institute of Science, Bangalore, India where he worked on developing Online Handwriting Recognition Systems for Tamil language. From 2015 to 2017, he was a Deputy Engineer at Bharat Electronics Limited (BEL), Bangalore, India. While at BEL, he was involved in the upgradation of Tropo-Communication System of the Indian Airforce (South Western Air Command). His research interests include machine learning, neural signal processing and behavioural neuroscience.

Mr. Jerrin was a recipient of Ministry of Human Resource and Development (MHRD) Research Fellowship in 2017 and Pratiksha Trust International Travel Grant in 2018.



RAMAKRISHNAN ANGARAI GANESAN (M'94–SM'95) is a Professor of Electrical Engineering at the Indian Institute of Science. He has graduated 32 research students and guided over 80 Masters theses. He formulated the problem of script recognition at the level of words in printed documents and proposed attention-feedback segmentation of the individual symbols from online handwritten words. He led a national research consortium on handwriting recognition in 8 languages and was one of the coordinators of the consortium on document image understanding in 12 scripts. Blind students are using over 600 Braille books in Tamil, converted from printed books using his Mozhi Vallaan OCR, which earned him the Manthan Award 2014 in the category e-inclusion and accessibility. He has developed unrestricted vocabulary, handwritten word recognition systems for Tamil and Kannada. He proposed a new algorithm for pitch synchronous pitch modification using DCT in the source domain. He has also developed Thirukkural and Madhura Vaachaka - TTS for Tamil and Kannada, used by blind students, for which he received the Manthan Award 2015 in the e-education category. He conceived of Linguistic Data Consortium for Indian Languages, currently managed by CIIL, Mysore. He is a member of the FICCI - Indian Language Internet Alliance. For his earlier work on nerve conduction in leprosy, he received the Sir Andrew Watt Kay Young Researchers Award from the Royal College of Physicians and Surgeons, Glasgow. He was the President of Biomedical Engineering Society of India and is a Fellow of the Indian National Academy of Engineering. He is an Associate Editor for Frontiers in Neuroscience - brain imaging methods.